Teng Jian Khoo (HU-Berlin/Innsbruck - ATLAS)
Paul Laycock (BNL - Belle II, DUNE)
Andrea Rizzi (INFN Pisa - CMS)

Scope & Activities

Outline

- Mandate & Goals
- Recent events (pre-COVID)
- HL-LHC computing review
  - Near- & medium-term targets
- Outlook

Working group page:
https://hepsoftwarefoundation.org/workinggroups/dataanalysis.html

DAWG Goals

Aims:

- Reduce monotonous and laborious tasks in physics analysis
- Optimise human and computing costs of publishing physics results

Priorities:

- Define problems by identifying the needs of physicists and the requirements of analyses across experiments via direct consultation
- Find solutions by connecting physics analysis experts and technological innovators within and beyond the HEP community

Highlight event

Pre-CHEP ‘19 WLCG/HSF Workshop

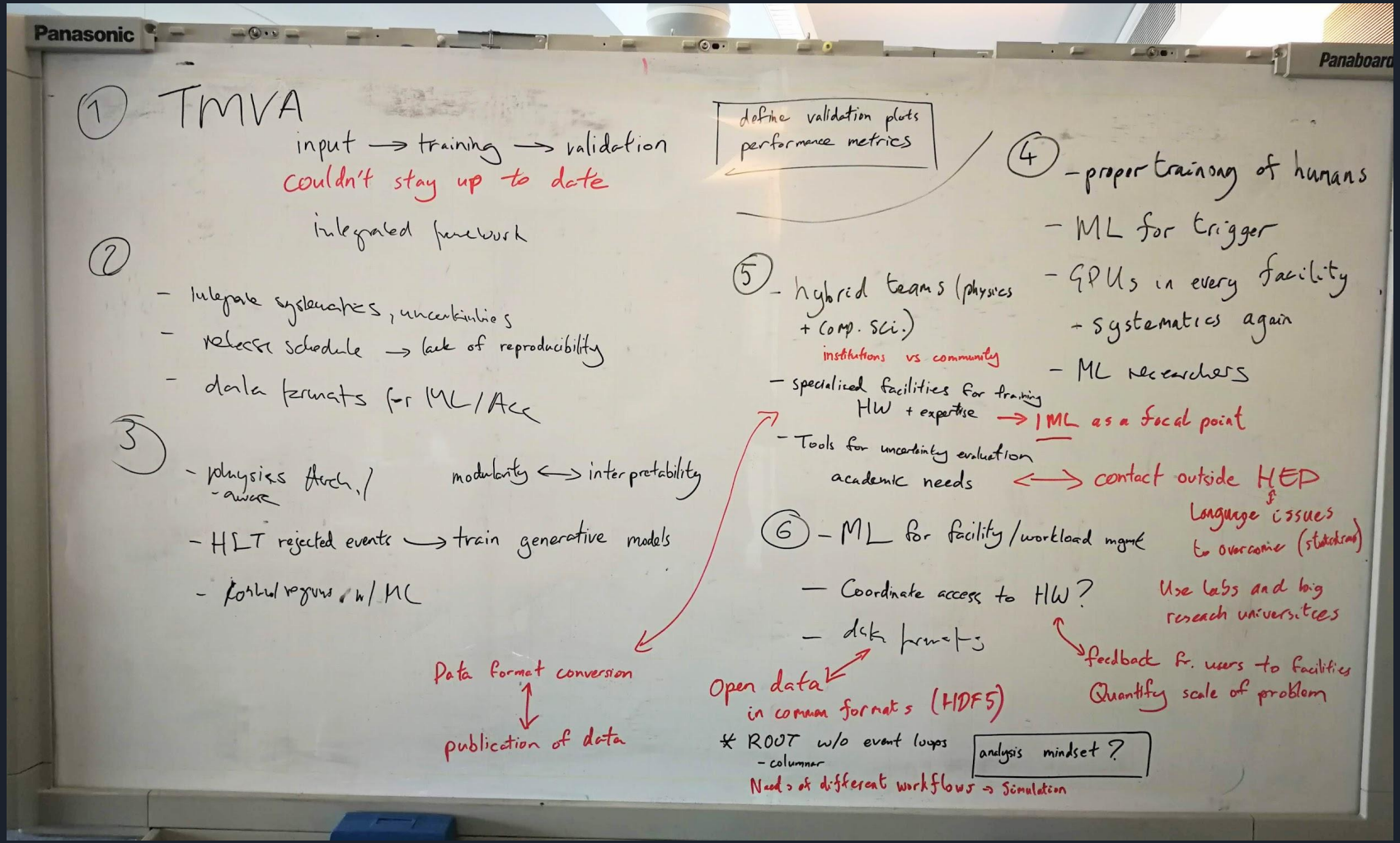
Analysis Systems: From Future Facilities to Final Plots

“Brain-writing” exercises addressing:

- Future analysis models
- Facility requirements for high-throughput analysis
- Growing integration of Machine Learning

Continued active engagement with WLCG critical

- Following up with DAWG/DOMA meetings on analysis facilities


Panasonic
Panaboard
1 TMVA
input → training → validation
couldn't stay up to date
integrated framework
define validation plots performance metrics
2
- integrate systematics, uncertainties
- release schedule → lack of reproducibility
- data formats for ML/Acc
3
- physics driven / -aware
modularity ↔ interpretability
- HLT rejected events → train generative models
- control regions w/ ML
4
- proper training of humans
- ML for trigger
- GPUs in every facility
- systematics again
- ML researchers
5
- hybrid teams (physics + comp. sci.)
institutions vs community
- specialized facilities for training HW + expertise → ML as a focal point
- Tools for uncertainty evaluation
academic needs ↔ contact outside HEP
Language issues to overcome (students)
6
- ML for facility/workload mgmt
- Coordinate access to HW?
Use labs and big research universities
- data formats
feedback fr. users to facilities
Quantify scale of problem
Data format conversion
publication of data
Open data
in common formats (HDF5)
* ROOT w/o event loops
- columnar
analysis mindset?
Needs of different workflows → Simulation

HL-LHC Computing Review

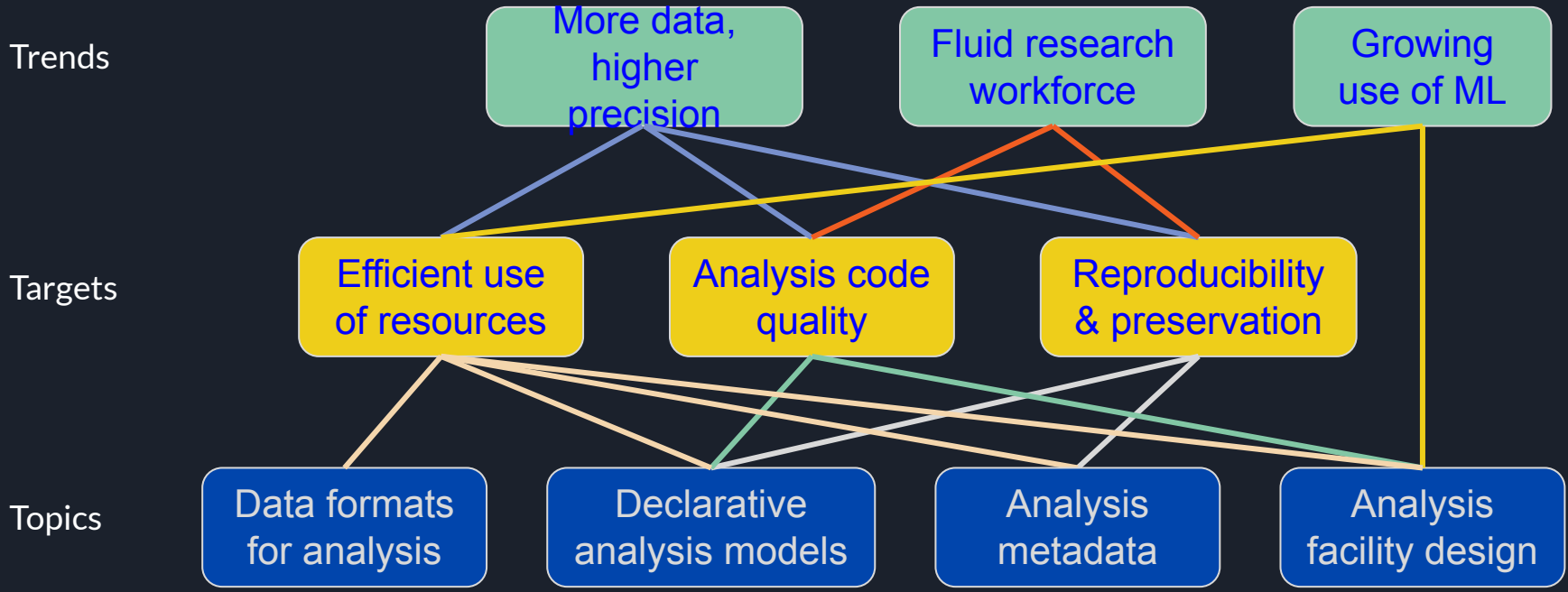
Update of Community White Paper

LHCC commissioned review by HSF: “Common Tools and Community Software”

Analysis highlights:

- Analysis data formats -- centralised production, disk costs, data access patterns, systematic uncertainties
- Metadata handling -- bookkeeping analysed data (does processing 100% of data scale to HL-LHC?), validity & retrieval of calibrations, cross-sections, …
- Quality assurance -- code testing for accuracy & efficiency
- Analysis interfaces -- declarative configuration, transparency, preservation

Development targets



Specific questions

Standardised analysis formats *a la* CMS nano-AOD, ATLAS DAOD_PHYSLITE
-- Production models? Adaptability c.f. the “10% analyses”

Analysis interfaces, description, preservation
-- Is a Domain-Specific Language a practical solution?
-- Or declarative layers (high-level workflow, mid-level tasks, low-level cuts)?
-- How to store/access metadata uniformly and robustly?

Analysis & the grid
-- What do we need at computing facilities (GPU, fast network vs disk, …)?
-- Do we need specialised facilities for analysis? How will job distribution work?
-- How to improve validation & performance monitoring of user code?

Outlook

Analysis software should be an enabler, not an obstacle
-- Design such that good practices are the default

Build capabilities for growing sophistication without exploding costs
-- Need effective interfaces to ML, accelerators
-- Must provide equitable access to infrastructure

Close connections to software training & documentation
-- “Higher level” languages for analysis operations could help

Quis custodiet analysis metadata?
-- Do we need an event/body to steer? Key stakeholders?